Technical Report 680

Forecasting Device Effectiveness:
Volume I. Issues

Andrew M. Rose and George R. Wheaton
American Institutes for Research

and

Louise G. Yates
Army Research Institute

Training and Simulation Technical Area
Training Research Laboratory

U. S. Army

Research Institute for the Behavioral and Social Sciences

June 1985

DISCLAIMER NOTICE

THIS DOCUMENT IS BEST QUALITY PRACTICABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

U. S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the

Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Technical Director

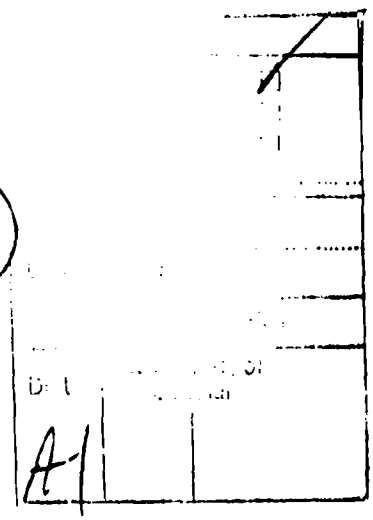
L. NEALE COSBY
Colonel, IN
Commander

Research performed under contract
for the Department of the Army

American Institutes for Research

Technical review by

Joseph D. Hagman
Michael J. Singer



NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-TST, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: ARI Technical Report 680	2. GOVT ACCESSION NO.: AD-A155 576
	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): FORECASTING DEVICE EFFECTIVENESS: VOLUME I. ISSUES	5. TYPE OF REPORT & PERIOD COVERED: Final Report: Vol. 1 of 3 August 1982-December 1984
	6. PERFORMING ORG. REPORT NUMBER: AIR 25400 TR
7. AUTHOR(s): Andrew M. Rose, George R. Wheaton (AIR), and Louise G. Yates (ARI)	8. CONTRACT OR GRANT NUMBER(s): MDA 903-82-C-0414
9. PERFORMING ORGANIZATION NAME AND ADDRESS: American Institutes for Research, 1055 Thomas Jefferson Street, NW, Washington, DC 20007	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 2Q263744A795, 6220, 3.4.1
11. CONTROLLING OFFICE NAME AND ADDRESS: U.S. Army Research Institute for the Behavioral and Social Sciences, 5001 Eisenhower Avenue, Alexandria, VA 22333-5600	12. REPORT DATE: June 1985
	13. NUMBER OF PAGES: 93
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): --	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

Louise G. Yates, Contracting Officer's Representative. Volume II of Forecasting Device Effectiveness numbered RP 85-25; Volume III numbered TR 681.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Training device effectiveness
Acquisition of skills
Transfer of training
Predicting device effectiveness

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This Technical Report discusses a number of issues that bear on the development of formal analytic methods for predicting the potential effectiveness of alternative training devices. The discussion encompasses theoretical, practical, and methodological issues uncovered during review of the literature and analysis of the problem.

With respect to theoretical issues, two fundamental sets of concerns are discussed. First, what is actually meant by the term "device —" (Continued)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

ARI Technical Report 680

20. (Continued)

effectiveness?" In this connection, the following issues are addressed: What is transfer of training? How is it measured? What are the pros and cons of its use as a measure of device effectiveness? What are the alternatives to transfer of training as measures of effectiveness? The second concern regards the classes and types of variables that hypothetically, at least, influence device effectiveness. In this discussion, the classes of variables that are identified are treated within a program evaluation framework. This structure is introduced to help organize and conceptualize the training system design and evaluation problem.

In the discussion of practical and methodological issues, several topics related to real-world constraints on developing and evaluating a training device effectiveness forecasting procedure are considered.

Technical Report 680

Forecasting Device Effectiveness:
Volume I. Issues

Andrew M. Rose and George R. Wheaton
American Institutes for Research

and

Louise G. Yates
Army Research Institute

Submitted by
Stanley F. Bolin, Acting Chief
Training and Simulation Technical Area

Approved as technically adequate
and submitted for publication by
Harold F. O'Neil, Jr., Director
Training Research Laboratory

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

June 1985

Army Project Number
2Q263744A795

Training and Simulation

ARI Research Reports and Technical Reports are intended for sponsors of R&D tasks and for other research and military agencies. Any findings ready for implementation at the time of publication are presented in the last part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

FOREWORD

Army training developers need tools to aid in the design, acquisition, and use of simulation- and computer-based programs of instruction for weapon operation and maintenance. One critical need is a job aid for the design and evaluation of training devices during all stages in the weapon acquisition cycle.

This series of three reports describes one approach to such aiding--a hybrid of decision analysis and mathematical modeling. The approach provides numerical estimates of device effectiveness which are based on expert ratings of trainee and task characteristics, functional and physical similarity between the proposed device and the operational equipment, and the instructional characteristics of the device. It is an analytic, computer-based technique--a menu-driven system--which can be used at any stage of training device design.

The product of this research can help training device procurers such as PM-TRADE and training developers in TRADOC make better documented decisions about training device design.

Edgar M. Johnson

EDGAR M. JOHNSON
Technical Director

ACKNOWLEDGMENTS

Many individuals contributed to the ideas contained in this report, either through their own publications or during extended conversations with the authors. We are indebted to several staff of the Simulation Systems Design Team at the U.S. Army Research Institute who shared their insights and experiences with us. We particularly wish to thank Dr. John A. Boldovici, the Project Monitor, for his contributions.

We also wish to acknowledge the contributions of Mr. David L. Winter and Ms. Mildred Jarvis of AIR's Systems Divisions. Their vast experience in designing and evaluating training devices kept our feet firmly anchored in the real world.

Finally, we wish to thank Dr. Robert M. Gagne who reviewed and provided valuable comments on an earlier version of this report.

Forecasting Device Effectiveness

EXECUTIVE SUMMARY

Requirement:

To develop a conceptual framework and methodology for predicting the effectiveness of a training device or simulator; to analyze and summarize training device evaluation issues including criteria of training effectiveness, variables that influence effectiveness, and constraints that affect device evaluation in either its empirical or rational form.

Procedure:

A literature review was conducted and the process of acquiring training devices within the Life Cycle System Management Model was analyzed. Theoretical and practical issues of training device design, development, and evaluation were investigated. Results were used to construct a conceptual framework within which to develop a procedure for predicting device effectiveness.

Findings:

Training device evaluation can be viewed within the more general context of a program evaluation rationale. This model consists of a network of hypotheses that relate program inputs and activities to a series of intermediate outcomes that also are logically linked. The model provides for multiple criteria of training effectiveness. These include skill acquisition, transfer of training, and efficiency of training and transfer. The model also provides for several different classes of variables that hypothetically may influence effectiveness. In both of these respects, the conceptual framework is superior to earlier models that have been more narrowly focused.

Utilization of Findings:

An analytic method for forecasting training device effectiveness can be developed from the conceptual framework described in this report. Such forecasts are of value during the device acquisition process when opportunities to conduct empirical research and evaluation are severely limited.

FORECASTING DEVICE EFFECTIVENESS:
VOLUME I. ISSUES

CONTENTS

LIST OF FIGURES AND TABLES

FORECASTING DEVICE EFFECTIVENESS:

I. ISSUES

1. Introduction

This report is submitted in partial fulfillment of Contract MDA 903-82-C-0414 between the U.S. Army Research Institute (ARI) and the American Institutes for Research (AIR). It is part of a programmatic effort to develop and analytically evaluate a model designed to forecast training device effectiveness. This report, the first of a series, discusses a number of issues that bear on the development of formal analytic methods for predicting the potential effectiveness of alternative device designs. The discussion encompasses theoretical, practical, and methodological issues uncovered during our review of the literature and analysis of the problem.

Background

The Army relies on training devices and simulators as indispensible components of performance-based training. Devices can be designed to incorporate instructional features that, for example, provide for control of

feedback, repetition of exercises, freeze and playback, and adaptive sequencing of instruction; these features are associated with specialized hardware and software that are not typically available on the parent equipment. Likewise, devices are often safer, more available, and cheaper to use than operational parent equipment.

To support the acquisition of cost-effective training devices, the Army has formalized a four-phase process that is linked to the Life Cycle System Management Model (LCSMM) of the parent material system (Carroll, Rhode, Skinner, Mulline, Friedman, & Franco, 1980; CORADCOM, 1980; Kinton, 1980; Kane, 1981). Kane and Holman (1982) provide an idealized description of the four phases of device acquisition and the corresponding hardware development cycles.

In each successive phase of acquisition, training device design decisions presumably are based on more detailed and precise information about the training requirement to be met, the physical and functional characteristics of the device needed to satisfy that requirement, the manner in which the device will be utilized, its effectiveness and its cost. The intent of the many steps in the formal acquisition process is to insure that the initial and often vague training concept is translated into

cost-effective training equipment that troops eventually interact with, at school or in the field. The great appeal of a highly structured acquisition process is that its many phases and steps are conceptually coherent, promising a procedure for systematically raising and then empirically resolving training device design issues.

In practice, however, unavoidable logistical demands in the training device acquisition process and the LCSMM that supports it make implementation in its idealized form impossible. As a consequence, the design of cost-effective training devices continues to be fraught with difficulty. For example, constraints in the acquisition schedule imposed by development of the parent system often preclude empirical evaluations during the design and development process; if such an evaluation is conducted, for example at Operational Test (OT) I or OT II, it is usually too late in the acquisition process to modify device design based on the evaluation results. As a necessary consequence, appraisals of a particular design or of competing design alternatives are primarily analytic.

However, for several reasons -- lack of reliable and valid analytic tools, paucity of applicable research, etc. -- formal analytic procedures are inadequate or

nonexistent. The bases on which device design decisions are made have not been clearly articulated, nor is it clear what types and levels of data are needed to support each decision. Thus, there is a need for analytic procedures, applicable during both early and later stages of device acquisition, that permit prediction of the potential effectiveness of alternative device designs.

To date, only a handful of analytic methods and models have been developed that attempt to evaluate or predict the effectiveness of training devices. Most of these have emerged from a program of research sponsored by ARI. The objective of these efforts has been to develop methods to forecast transfer of training based on information about training device characteristics. There have been several recent reviews of these methods (e.g., Tufano & Evans, 1982; Harris & Ford, 1983; Knerr, Nadler, & Dowell, 1983). We will not repeat these reviews here; rather, we will summarize the limitations that one or more of these reviews have remarked upon.

- **None of the methods has been satisfactorily validated empirically:**
  - -- Virtually no empirical studies have been attempted;

-- A "criterion problem" of what to measure and how to measure performance has limited the evaluation of the methods;

-- In many cases, it is not feasible to measure operational performance on the parent equipment.

- **The models have too narrow a focus:**
  - -- Extra-device variables (e.g., utilization, student and instructor acceptance, student capabilities, etc.) have not been included;
  - -- Device and system characteristics affecting learning have not been considered;
  - -- Models have not addressed such issues as criticality or importance of training.

- **The models have been inefficient to apply:**
  - -- The few that have been developed consist of tedious, manual, paper-and-pencil procedures;
  - -- They provide a microscopic level of analysis.

- **The models are of limited diagnostic utility:**

-- They arbitrarily aggregate judgmental data, thereby producing relatively uninterpretable summary indexes;
-- Algorithms and rationales for decisions based on obtained indexes are arbitrary or not specified.

Recognizing these limitations, ARI has sponsored the current project, the major objective of which is to build upon previous efforts and overcome their shortcomings. In support of this effort, AIR reviewed literature and conducted conceptual analyses to examine the utility of transfer as a dependent/criterion variable, explored alternatives and supplements to transfer for assessing device effectiveness, and ascertained variables hypothetically affecting various effectiveness criteria. Based on our findings, we provided recommendations for alternative or supplemental criterion measures, for modifications of ARI's ADP-based effectiveness forecast system, and for additional research.

Organization of This Report

This report is organized around several issues related to the evaluation of effectiveness. For each major issue, we address a number of questions, present various arguments, and attempt some resolutions where appropriate.

In the following chapter, we discuss two fundamental theoretical issues. First, what actually do we mean by the term "device effectiveness?" That is, what should be the criterion of device effectiveness and how should it be measured? In this latter connection, we address the following questions: What is transfer of training? How is it measured? What are the pros and cons of its use as a measure of device effectiveness? What are the alternatives to transfer of training as measures of effectiveness? In this regard we discuss several possibilities, including acquisition of skills and knowledge, acquisition efficiency, and other concepts.

The second major issue concerns the "content" of an effectiveness evaluation model: What are the classes and types of variables that hypothetically, at least, influence device effectiveness? In this discussion, we introduce a "program evaluation framework" to help organize these variables and to aid the conceptualization of the training system design and evaluation problem.

In Chapter 3, we discuss practical and methodological issues related to real-world constraints on developing and evaluating a training system effectiveness forecasting procedure. Topics include the impact of the LCSMM,

difficulties of criterion measurement, constraints on statistical techniques used in evaluations, and limitations on the measurement of variables.

2. Theoretical Issues

Overview

An ideal methodology for analytically evaluating (or forecasting) the effectiveness of a training device or simulator would have several properties. First, in accord with the existing LCSMM, it would be applicable at different stages of device design and development. Second, it would be diagnostic -- it would indicate which device features contributed to effectiveness and which ones detracted from it. Third, it would be easy to use. Fourth, it would support different levels and types of decisions (e.g., "Will Device 1 shorten skill acquisition time on the operational equipment?" "Is Device 1 more cost-effective than the alternative designs?").

When contemplating development of a method for evaluating devices one immediately encounters two fundamental sets of concerns. First, what actually do we mean when we say that a device is "effective?" What would be our criterion of device effectiveness and how would we measure it? Second, what would be the content of our forecasting method? What are the classes and types of variables that would (or could) influence device effectiveness? These two

concerns -- specification of criterion dimensions and specification of predictor variables -- are addressed in this chapter.

Issue: What is Device Effectiveness?

What do we mean when we claim a device is "effective?" Traditionally, effectiveness is usually expressed in in terms of transfer of training. We will discuss this concept below. Following this discussion, we will present other potential criteria of effectiveness.

Transfer of training: Definition. "Transfer" has been used to refer to an <u>empirical</u> phenomenon, defined by the results from specific experimental paradigms. For example, a simple transfer paradigm is:

Group 1: Trains on Training Device A -->
Trains to criterion performance on operational task

Group 2: No training -->
Trains to criterion performance on operational task

To the extent that Group 1 reaches operational proficiency faster than Group 2, we say that Group 1 has benefited by "positive transfer." Thus, transfer is defined as the beneficial (or harmful) effect of specific

previous learning on the learning of a new task. Depending on the paradigm and the measures of performance used, we can define "first-trial" transfer (i.e., the beneficial or harmful effect of specific previous learning on initial performance of a task), "long-term" transfer (the effect of previous experience on the rate of skill acquisition on a new task), and other transfer terms. The important point is that "transfer" is defined by the experimental paradigm and measure of performance used; it is an index of differential performance produced by specific experimental manipulations. (For a further discussion of transfer indexes and theoretical underpinnings, see Appendix A).

Transfer has been the principal criterion of training device effectiveness in most previous attempts to develop methods for predicting device effectiveness, including all of the TRAINVICE series (Wheaton, Fingerman, Rose, & Leonard, 1976a; Wheaton, Rose, Fingerman, Korotkin, & Holding, 1976b; Hirshfeld & Kochevar, 1979; Narva, 1979a, 1979b; Swezey & Evans, 1980; Faust, Swezey, & Unger, 1980). The rationale for transfer as the criterion is straightforward: Device 1 is more effective than Device 2 if, after completing training on each device, trainees who used Device 1 perform better (i.e., initial transfer) or achieve proficiency faster (i.e., rate of skill acquisition) on the operational task, than trainees who used Device 2.

Transfer of training: Limitations. There are two important criticisms of this "transfer" rationale for device evaluation. First, some form of operational performance must be measured. This calls for an elaborate specification of "criterion performance," including such considerations as allowable individual variation, control for measurement error, alternative performance measures, etc. Obviously, the more complex the operational task, the more difficult such specifications are to elaborate. For something complex like "Hit a moving target" in tank gunnery, such elaborations rapidly become arbitrary (e.g., which of myriad conditions should be tested? How reliable is the weapon? Is a test on a controlled range at Fort Knox, using targets that don't shoot back, an adequate surrogate of "actual" combat? etc.). However, for many other tasks, the specifications are much more straightforward (e.g., convert grid to magnetic azimuths; change the brake linings on a jeep). More simply, there is a continuum of operational task complexity that is reflected by criterion measurement problems.[1] Having chosen transfer as a criterion of device effectiveness, one must be prepared to deal with these measurement problems. Adequate measurement

---

[1] We discuss the practical issues of criterion testing in a later section, where we also indicate how one would validate a model that predicts transfer.

of operational performance may often be difficult or, in extreme cases, impossible. But this prospect should not lead to the rejection of transfer as a criterion of device effectiveness; if performance measurement is impossible, surrogate measures of transfer could still be considered.

The second major criticism of the transfer rationale is that it is too restrictive: it ignores the time, cost, and effort associated with the actual accomplishment of training.[2] To use an extreme example, suppose two devices demonstrate the same amount of transfer; however, trainees on Device 1 must spend ten times longer practicing on it than on Device 2. Clearly, these devices are not equally effective except in the most general (transfer) sense.

Another way of stating this criticism is to argue that a training device could and should be viewed as part of the larger training program in which it is embedded: a device

---

[2] Traditionally, the "goodness" of any training system is expressed along two dimensions: cost and effectiveness. In addition to direct acquisition and production dollars, "cost" has several other components that, in the training device situation, are convertible to dollars. Device facility requirements, student throughput, student-to-instructor ratios, repair and replacement time, device reliability, and other standard cost components fall into this category. While these components can (hypothetically) and should be dealt with systematically, they are not within the scope of this current effort. Nevertheless, we do treat general cost concepts as part of an overall training system evaluation approach.

is effective if it reduces the total time, cost, and effort needed to bring soldiers to operational readiness on the parent equipment. This more global view is in contrast to the narrower transfer rationale, which views device effectiveness solely in terms of the proficiency levels observed on the parent equipment. We will expand upon this point in a later section.

Transfer: Conclusion. From a common-sense perspective, the transfer rationale is unarguable: unless use of a training device promotes some positive benefit for operational performance (a savings in time to reach criterion proficiency, better first-trial performance, or whatever), it cannot be considered "effective." Thus, positive transfer, if the appropriate empirical evaluation could be conducted, would appear to be a necessary condition for a training device to be judged effective.

But, positive transfer, even when it can be assessed empirically, surely is not the only characteristic of an effective training device; total training time, cost, and effort must also be considered.

Other effectiveness criteria. If device evaluators (or purchasers) were told that two devices produced equal transfer scores (or that it was impossible to measure

operational performance), what else would they want to know about the devices? The evaluators might want to know what the trainee learns (or is supposed to learn) on each training device and its relevance to the operational task. In the example above, perhaps the extra time associated with Device 1 is due to training more knowledge and skills than is possible with Device 2 or even to training irrelevant knowledge and skills. The evaluators also might want to know if what is taught is taught efficiently. Similarly, they also might inquire about the efficiency with which the device prepares the trainee for the operational task. Both "acquisition efficiency" and "transfer efficiency" would entail an examination of the device's instructional features. One can think of other kinds of information that the evaluators also would like to have. Each of these additional types of information is considered below as a potential component of a criterion measure of device effectiveness.

Other effectiveness criteria: Acquisition of skills and knowledge. During the training device acquisition process, device evaluators may face two types of problems: first is the case where it is infeasible or impossible to obtain training or transfer data. Second is the case where empirical transfer-of-training evaluations are conducted

but the alternative devices do not differ on transfer index values. In the former case, evaluators would have to develop a surrogate measure or an estimate of "potential" transfer. In the latter case they would have to develop different measures or estimates of effectiveness. In both cases, the evaluators could expand their appraisal to look at the content of training: what is taught and how efficiently it is taught.

The "what" of training, when viewed as a surrogate measure of transfer, is typically measured as the degree of overlap between the content of the training objective and the operational performance objective. An index based on such overlap would represent the amount of required knowledge and skills the trainee has learned (or conversely, still must learn when the trainee progresses to the parent equipment).

Concepts regarding the content and overlap of training are usually derived from the various theoretical views of transfer phenomena. (See Appendix A for further elaboration of these theoretical views.) For example, based on Thorndikean "identical elements," one could look for specific high-fidelity simulations or duplications of the parent equipment and task(s) in the training device. In

the extreme, those adopting this view might argue that the effectiveness of training (and the criterion measure of device effectiveness) depends exclusively upon the number or percentage of these identical elements. According to this view, if one is to maximize effectiveness one must build the device to simulate the parent equipment to the maximum extent possible; i.e., a high fidelity simulation is required in which the content of training almost perfectly overlaps with that of the operational performance objective. And, of course, many devices are designed and developed with precisely this view in mind.

The "Osgoodian" view considers stimuli and responses along a continuum of similarity. Thus, the relevant content of training would be the stimuli and responses common to both situations, weighted somehow by their degree of similarity. An Osgoodian also might assert that a device that was identical in all respects to the parent equipment would be maximally effective. But he would allow for degrees of similarity in overlapping content, and would be able to generate predictions of different "degrees" of transfer; further, based upon an inspection of the content of training he would be able to predict the circumstances leading to <u>negative</u> transfer.

However, neither of these theoretical perspectives on the content of training addresses another commonly used training concept -- namely, enabling skills or knowledges. These are "things" that are necessary for operational performance but are not themselves directly a part of the criterion performance. More generally, an enabling skill or knowledge, once learned, increases the speed or efficiency of the learning of some other skill. Gagne (1965), for example, writes about hierarchies of skills and knowledges, where lower-order skills are necessary to learn higher-order ones, which are necessary for still higher-orders, and so on. In essence, one must learn to walk before one can learn to run. There need be no "identical elements" nor "stimulus-response similarities" at all between the lower-order enabling skills acquired in the training device and the higher-order skills comprising operational task performance on the parent equipment.

Many devices and training systems are designed and developed to teach enabling skills. "General maintenance trainers" are a good example: they are designed to teach prerequisite knowledges and skills that will enable trainees to acquire system-specific skills more easily. The important point is that the content of training cannot be delineated in terms of "identical elements" or

"stimulus-response similarities." The most suitable vocabulary to describe this type of training content is that used by cognitive psychologists (e.g., Neisser, 1976), who talk of "knowledge structures" and "schemas." Training consists of the building of an organized knowledge structure about a topic. This structure has "slots" where new information can be added to it. Thus, the goal of training is to develop knowledge structures in trainees that will enable them to incorporate new information -- the operational task -- easily.

Regardless of one's perspective or vocabulary, it is clear that an assessment of the content and relevance of the training device is, or should be, part of the characterization of a device's effectiveness. Content specification in terms of the device-mediated learning objective is obviously critical to the device designer/developer; it is also important to the training program evaluator in that it could serve as a surrogate measure when it is infeasible or impossible to obtain an empirical assessment of transfer.

Other effectiveness criteria: Acquisition efficiency. Suppose we have two devices, both producing the same "amount" of transfer and/or both teaching the same content. However, a trainee on one device takes ten times as long to

reach proficiency on that device (i.e., to acquire the content) as it does a trainee on the other device. Clearly, when everything else is equal, we would call the device that promoted more rapid learning the more "effective" one. The concept here is "efficiency": how well (rapidly, cheaply) does the device train the required content?

The "efficiency" of training typically is measured in terms of the rate of acquisition of the training objective. The resulting index would represent the time, cost, or effort required to reach proficiency on the training device.

Some aspects of the evaluation of efficiency include an examination of the device's instructional features and its pattern of use. For example, several training experts (e.g., Braby, Henry, Parris, & Swope, 1975) have developed prescriptive methods for the design of training based on analyses of instructional features. Typically, the form of the argument is, "In order to teach task type X effectively, a device must have feature Y." These arguments are then combined to produce preliminary device specifications. Clearly, it is a relatively straightforward matter to turn this argument around to generate evaluative criteria for assessing device effectiveness. Thus, "Device 1 has

feature Y; therefore, it will teach task type X effectively." If X is what we want to teach, Device 1 will be a more effective device than Device 2, which does not have feature Y.

However, care must be taken when examining instructional features, in that "more" does not necessarily imply "better." Devices with video playback and freeze-frame capabilities are not always better than devices without them (Swezey, Criswell, Huggins, Hays, & Allen, 1985). The effectiveness of a given feature will vary as a function of the training content. Much of the empirical research in this area uses "task type" as the descriptive vocabulary for training content (Braby, et al., 1975; Wheaton, et al., 1976a).

Other effectiveness criteria: Transfer efficiency. Suppose that two devices train the same content, and do so equally efficiently. They will not necessarily produce the same amount of transfer. This fact gives rise to another potential component of device effectiveness -- namely the efficiency with which the trainee is prepared for acquiring the skills and knowledges that still must be learned on the parent equipment. Instructional features can be incorporated in a device that enhance the rate of

acquisition of knowledge and skills on the parent equipment independently of enhancing the rate of acquisition of the device-mediated training objective.

A further fairly subtle point is that features that enhance transfer may not necessarily enhance acquisition. Suppose a training device had a feature that allowed for simulation of environmental conditions found in the operational situation -- noise, heat, darkness, etc. This feature would undoubtedly enhance transfer to these situations. However, its use would surely slow down the rate of skill acquisition or learning within the device.

Thus, transfer efficiency seems to be another distinct component of device effectiveness, in addition to those previously discussed: transfer, the content of training, and the efficiency of training. Are there other concepts that have been used or suggested as device effectiveness measures?

Other effectiveness concepts. Most other concepts that have been considered as potential measures of device effectiveness fall into the category of "user acceptance" (Mackie, Kelly, Moe, & Mecherikoff,, 1972). This usually has two parts: instructor acceptance and trainee acceptance. A device presumably will not be effective if

instructors and trainees won't or can't use it. Such might be the case, for example, if there were a significant burden added to instructors' workloads by requiring them to learn to operate a complicated device, if trainees had to learn excessive "extra-job" skills just to operate a device, or if either group felt the device was providing irrelevant training.

These are important considerations, certainly. A device should not be built or purchased that is too difficult or awkward for instructors and trainees to use. Presumably, indexes of instructor and trainee workloads could be incorporated in an assessment of device effectiveness. "Extra-job" skills could be incorporated as part of an index of the content and relevance of training. On the other hand, beyond emphasizing sound human-engineering practices (e.g., Smode, 1972), there is little that can be done by the device designer to increase the probability that the device will be considered relevant to instructors and trainees. Some might argue that acceptance will increase if the device can be made more realistic -- in other words, to make it simpler to relate the training to actual job performance. However, increased realism might or might not lead to more effective training, especially given the arguments made above concerning enabling skills. The real

issue is how best to convince instructors and trainees that the training system will lead to better job performance. In our opinion, the best way to do this is by providing them with empirical evidence of successful training.

Summary: Device effectiveness. The first step in developing an analytic procedure for predicting the potential effectiveness of training devices is to pin down just what we mean by the term "device effectiveness." In the preceding section we have examined several different and general conceptions of effectiveness: 1) an effective device promotes transfer of training to the parent equipment; 2) an effective device enables trainees to acquire necessary skills and knowledge rapidly; 3) an effective device is accepted by the trainees and instructors who interact with it.

The criterion most often used to characterize training device effectiveness is transfer of training, based on an estimate of trainee proficiency on the parent equipment relative to the proficiency of some type of control group on that same equipment. As we indicated earlier, when the estimate is based on an empirical investigation, transfer can be expressed in several different ways depending upon the specific experimental paradigm employed. For example,

relative to the performance of a particular type of control group, device effectiveness can be stated in terms of the level of trainee proficiency on the parent equipment after a specified amount of time (or trials) and/or as the amount of time (trials) required to reach a specified level of proficiency.

A second component or criterion of device effectiveness is the skills and knowledge acquired during training, expressed as an estimate of trainee proficiency on the training device per se. When based upon an empirical assessment, this estimate also can be expressed in different ways. For example, effectiveness can be characterized in terms of the level of trainee proficiency on the device after a fixed amount of practice (time, trials) or as the amount of practice required to attain a specified level of proficiency. In this connection, we noted that aspects of training external to and apart from the device (e.g., courses and lessons, classroom exercises, other training devices, etc.) may nevertheless contribute to proficiency on the device.

A third component of device effectiveness is user acceptance. This concept is typically operationalized in terms of trainee and instructor ratings. The ratings are

obtained on such training device dimensions as fidelity or realism, convenience of use, and the perceived value of training.

Although we have treated these notions of device effectiveness separately, we do not mean to imply that they are necessarily independent, alternative, or competitive criteria. Rather, we view them as useful and complementary components of an effectiveness criterion that is inherently multidimensional. To support the evaluation of a training device we would like empirical assessments of each component, whenever possible. While it may be highly desirable to determine how much transfer is associated with a given device, such a determination may not be feasible; or if feasible may be inconclusive; or when conclusive, may not tell the whole story. For these reasons, the empirical evaluation of a training device should encompass consideration of other components as well. Similarly, procedures for forecasting device effectiveness, which heretofore have focused entirely on transfer of training, also need to adopt this broader perspective.

This brings us to one of the most fundamental issues in this paper. How are we to proceed with the evaluation of a training device when the various components of device

effectiveness can not be assessed empirically, the situation typically confronting the designers and developers of major training devices? The answer lies in identifying surrogates for the components of device effectiveness discussed above, and then using analytic procedures to generate estimates of the various surrogates. For example, it might be possible to use amount of overlap in the content of training and operational (i.e., parent equipment) performance objectives as an estimate of potential transfer of training. Similarly, analyses of the content of training and performance objectives, coupled with an appraisal of instructional features, might provide estimates of acquisition or transfer efficiency. One objective of the present project is to identify such surrogates and to develop procedures for their assessment.

Issue: What are the Variables Influencing Device Effectiveness?

During the design, development and evaluation of training devices we need to consider the independent variables hypothetically influencing device effectiveness for two important reasons. First, when we are able to carry out an empirical evaluation of a training device, we will wind up with a multidimensional assessment that is almost

entirely outcome oriented. That is, we will describe the device in terms of a certain amount of transfer, a particular rate of skill and knowledge acquisition, etc. If at all possible, it would be desirable to augment such an appraisal with more diagnostic information that suggests how particular independent variables contribute to measured effectiveness. Armed with such knowledge, it would then be possible to entertain "what if" questions, contemplating in at least a rough fashion how device effectiveness might vary were changes in selected independent variables introduced. In this application, information about the relationships between independent variables and effectiveness criteria would be used to prescribe design modifications intended to enhance device effectiveness.

The second reason that independent variables hypothetically influencing device effectiveness are of interest is because an empirical evaluation of effectiveness often may not be feasible. In this case we would want to conduct an analytic appraisal and would need a set of predictor variables in terms of which to couch our effectiveness forecasts or estimates. That is, given information about selected independent variables, we would attempt to predict training device effectiveness on a variety of surrogate criterion measures. There also, of course, is

diagnostic value in such an appraisal. In principle, we could explore the manipulation of specific independent variables, estimating their influence on effectiveness, and use the results of various changes to inform us about the probable value of different design modifications.

Given a multidimensional criterion of device effectiveness that includes facets of both initial learning and subsequent transfer, we can think of many variables that potentially may influence device effectiveness, and therefore should be considered for diagnostic and forecasting purposes. Reviews of the literature and analyses of training phenomena (e.g., Miller, 1954; Valverde, 1968; Blaiwes, & Regan, 1970; Blaiwes, Puig, & Regan, 1973; Aagard & Braby, 1976; Wheaton, Rose, Fingerman, Korotkin, & Holding, 1976b; Royer, 1979; Hays, 1980; Rose, 1980; Rose, Allen, & Johnson, 1982; Rose, McLaughlin, & Felker, 1981) point toward a myriad of relevant variables for which there is empirical or theoretical support.

Based upon a review of the literature, an examination of available effectiveness forecasting models, and a multidimensional conception of training device effectiveness, there appear to be five categories of independent predictor variables that warrant consideration. That is, these

categories appear salient. If we were to manipulate variables within any of these categories we would expect to observe certain specifiable changes in particular components of the device effectiveness criterion. We discuss each category briefly.

Trainee quality. As the primary input to the training process, we are concerned about a variety of trainee variables. These include such concepts as trainee intelligence, aptitude or ability, motivation to learn, and prior experience, as reflected in entry levels of skill and knowledge and initial levels of proficiency on the training device or the parent equipment. Collectively, such variables represent the quality of incoming trainees and are usually manipulated as part of some earlier personnel selection or classification procedure. It is hypothesized that higher quality will be reflected in faster rates of skill acquisition and greater or more rapid transfer.

In many contexts, personnel variables of this type are treated as within-group individual differences, with a focus on each individual. Traditionally, however, training device designers and evaluators have addressed quality of personnel essentially as a between-group variable. That is, device developers have predicated certain design

decisions on the characteristics of the typical, average, or modal trainee who will proceed through training. Device evaluators have attempted to match experimental (trained) and control (untrained) groups on the basis of trainee quality during empirical assessments of transfer of training.

Preliminary training. Variables within this category reflect the type and amount of enabling or prerequisite instruction and training that trainees receive prior to their exposure to the training device. Indoctrination and orientation sessions, procedural training, demonstrations, lectures and reading assignments, etc., that enhance the quality of trainees and better prepare them for device-mediated training fall within this category. It is hypothesized that the provision of enabling skills and knowledge, proficiency in part-task performance, etc., will be associated with more rapid acquisition of training device-mediated objectives and better (greater, faster) transfer.

Task type. The types of tasks comprising a device-mediated training objective or the operational performance objective associated with the parent equipment are important considerations. The type of task includes such

variables as the number of task steps, sequential dependencies among steps, task aiding, cognitive and psychomotor demands, etc. Systematic manipulation of these types of variables is known to influence acquisition and retention of skilled performance and should influence acquisition and transfer components of device effectiveness.

Device type. This category includes variables that represent engineering and instructional features of a training device. These features are the ones that typically come to mind when designers and evaluators ponder about characteristics that may enhance or degrade training device effectiveness.

The subset of so-called <u>engineering variables</u> reflects such concepts as the fidelity of simulation or similarity between the training device and the parent equipment it presumably represents. In spite of a voluminous literature on concepts like engineering, environmental, or psychological fidelity, or physical and functional similarity, their influence on components of device effectiveness is not clearly understood. Very generally speaking, increases in similarity between the device and parent equipment facilitate transfer of training. However, very high similarity or fidelity does not insure better transfer;

transfer of training can occur when fidelity, at least as conventionally measured, is quite low; and there are conditions of stimulus and response similarity that can lead to at least initial if not prolonged negative transfer of training.

The subset of instructional features includes variables that are intended both to facilitate acquisition of skill in the training device and to promote transfer of training to the parent equipment. These variables include sequencing of stimulus or problem difficulty, provision of feedback to both trainees and instructors, manipulation of signal-to-noise ratios, measurement and recording of trainee performance, adaptation of type and level of instruction to level of proficiency, etc.

Training context. This category subsumes a variety of ancillary but potentially important variables that do not fit neatly into any of the prior categories. The variables are descriptive in one way or another of the larger training program or context within which a training device is utilized. For example, contextual variables include the scheduling of training (e.g., the type, amount and distribution of practice) as well as the performance criteria that signal a cessation of training on the device and

adequate proficiency on the parent equipment (e.g., first-trial or longer-term transfer). They also include instructor proficiency as well as user acceptance of the device.[1]

All of the variables subsumed under these categories are familiar. The issue is, which ones of this large array need to be considered, particularly in the course of developing a procedure to forecast training device effectiveness? In general, existing methods have focused almost exclusively on training device parameters, choosing largely to ignore extra-device, training program variables. Two rationales have been advanced for this restricted focus. The first is that forecasting procedures do not want to "penalize" a device -- e.g., with a lower effectiveness score -- simply because it might be used inappropriately, introduced without prerequisite instruction if required, or staffed and operated by poorly trained instructors, etc. The second and more pragmatic reason is that information about the training program or device utilization is seldom supplied along with a detailed description of the training

---

[1] User acceptance, as our earlier discussion suggests, can be viewed as a criterion of device effectiveness. Our preference, however, is to treat it as an intervening variable. User acceptance, therefore, can exert an influence on the primary acquisition and transfer components of device effectiveness.

device. At best, therefore, present forecasting methods "reward" a device that allows for flexibility of utilization, but do not provide for evaluation of the device in terms of a specific utilization plan or training program context. Below, we describe a general program evaluation framework that can be used to organize the effectiveness criterion and predictor variables discussed so far.

Theoretical Issues: Conclusion. A Device Effectiveness Evaluation Framework

Throughout the discussion of criterion and predictor variables of device effectiveness we have found it useful to broaden our perspective on device evaluation: to consider criteria of effectiveness in addition to transfer of training; to examine predictor variables lying beyond the domains of task and device characteristics that traditionally have been examined during empirical and analytic assessments of effectiveness. We believe that a training device, no matter how simple (e.g., a part-task trainer) or sophisticated (e.g., a full-scale weapon system simulator) is but one component of a larger training program. It is possible to compare training devices or even alternative training concepts that are in some sense interchangeable

within a given training program, but it does not make much sense to compare or evaluate them in the absence of such a broader context.

Given this larger perspective, it follows that a training device can not be meaningfully evaluated without considering its intended role in the overall program, including the plan for its use. Thus, what needs to be evaluated or compared is not the training device(s), but the entire training program(s). This includes the specification of training materials (documentation, devices, and instructors), the sequence of training or the program of instruction, the level of instructor training required and provided, the amount of instructor and student time involved, and the criteria for successful completion of the training program and operational proficiency on the parent equipment.

How does one evaluate an entire training program? In other words, given certain inputs (knowledges, skills, abilities, and other characteristics of the trainee population) and certain desired outputs (proficiency requirements of the operational situation), how do we evaluate the program that is designed to operate on the input to achieve the desired outcome?

Ultimately, we can express program effectiveness in terms of the extent to which terminal program objectives are met. Those objectives are to get trainees to criterion levels of operational proficiency as quickly, cheaply, and safely as possible. However, it often is infeasible or impossible to determine whether terminal program objectives have been met. Moreover, by focusing exclusively on terminal outcomes, one may neglect several other important evaluative criteria of the types discussed earlier that provide valuable diagnostic information -- why the program was effective or not effective.

Evaluation issues of these types have abounded in many other contexts, most notably during attempts to evaluate the impact of major social programs (e.g., Cronin & Bourque, 1981; Cronin, Drury, & Gragg, 1983). Although these programs (e.g., criminal justice, education, poverty, health care delivery, etc.) and the specific indexes of program impact developed for them have no bearing on training device evaluation, the basic model of impact assessment that has been employed is directly relevant: frequently, it was infeasible or impossible to measure terminal program objectives directly; diagnostic information was critical to the evaluation; there were many "extraneous" (to the program) variables that affected the outcomes.

As shown in Figure 1, the model is based on a program rationale, or network of hypotheses, which makes explicit the dynamics of the cause-effect relationships being investigated.

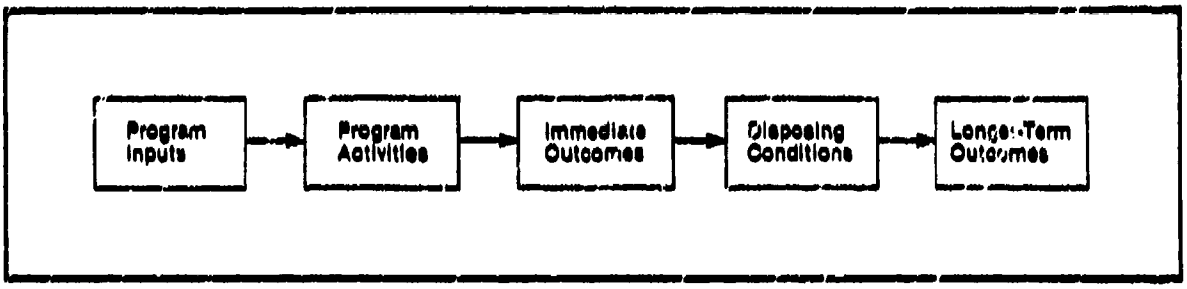


Figure 1. General model of the program rationale.

The methodological focus in this model is on the hypotheses that relate events at one stage to those at the next. The certainty with which outcomes can be attributed to inputs under program control is vastly enhanced by this technique. An important consequence of this feature is that the assessment does not treat an intervention program as an entity that succeeds or fails in accordance with the average impact yielded by the type of approach which characterizes the program. The aim is to identify the individual components that should be modified or attended to when further implementation or evaluation is planned.

This general type of program evaluation model seems perfectly suited to the assessment of training devices. It suggests that we examine the training program rationale: the specific cause and effect linkages that explain why and how certain inputs (planned and unplanned) lead to certain outcomes. Development and analysis of the rationale require description of many aspects of the training program, including: the input and ultimate output, all of the intermediate outcomes, the linkage between intermediate outcomes, the variables potentially influencing each intermediate outcome, and the relationships between the intermediate outcomes and ultimate program output.

An example of a rationale that links independent predictor variables to various components of training device effectiveness might look something like the following:

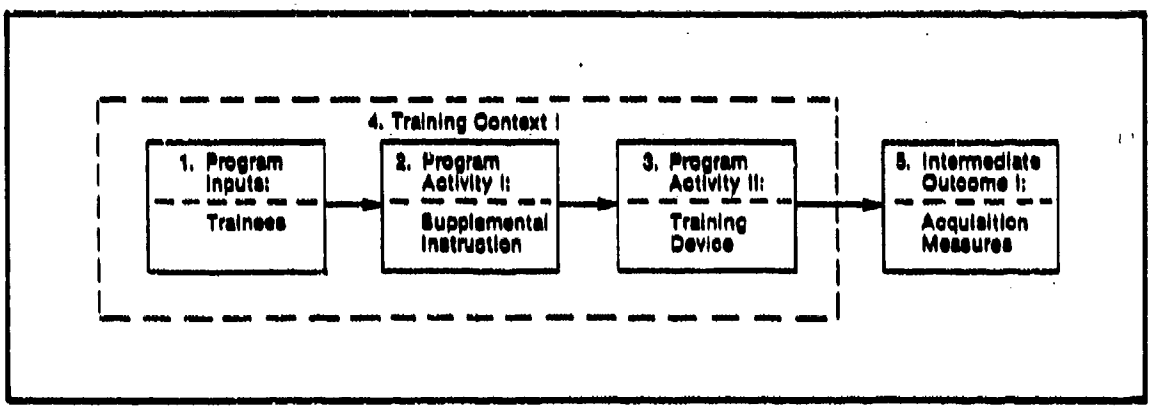
1. Program inputs are the learning-relevant characteristics of the trainees. These may be knowledges, skills, abilities and other characteristics including trainee motivation to learn. We have already mentioned such variables under the general rubric of trainee quality, a class of variables that can be manipulated to influence estimates of device effectiveness.

2. Program activity I is the preliminary training and instruction that trainees receive as part of the overall training program, prior to their practicing on the training device. Training programs obviously can differ widely in the amount and type of such support.

3. Program activity II is the training mediated by the training device per se. Its description would include the specific training objective(s), the types of tasks contained in the device-mediated training objective, and the instructional features with which the device is equipped. Physical and functional similarity as well as various types of fidelity would also be included as part of the training device description.

4. Training context I includes everything that potentially might affect the trainee-device interaction above and beyond the program elements already described. The context could include instructor proficiency, user acceptance, device reliability and maintainability, practice schedules, integrity (with respect to some plan) of device implementation, and interactions among these and other variables.

The training device evaluation model so far is:



5. **Intermediate outcome I** is trainee performance on the training device. This first component of device effectiveness can be expressed in terms of both time and accuracy measures of performance and in terms of "process" information (e.g., time, trials, acquisition rate, etc.). The focus is on the skills and knowledge that are imparted through device-mediated training as well as on the efficiency with which the training objective is accomplished. If trainee proficiency on the device does not reach expected levels, then we would perform diagnostic analyses to seek the reasons for such a shortcoming. Toward that end we would examine the trainee input, the supplemental instruction, characteristics of the training device, and facets of the larger program context.

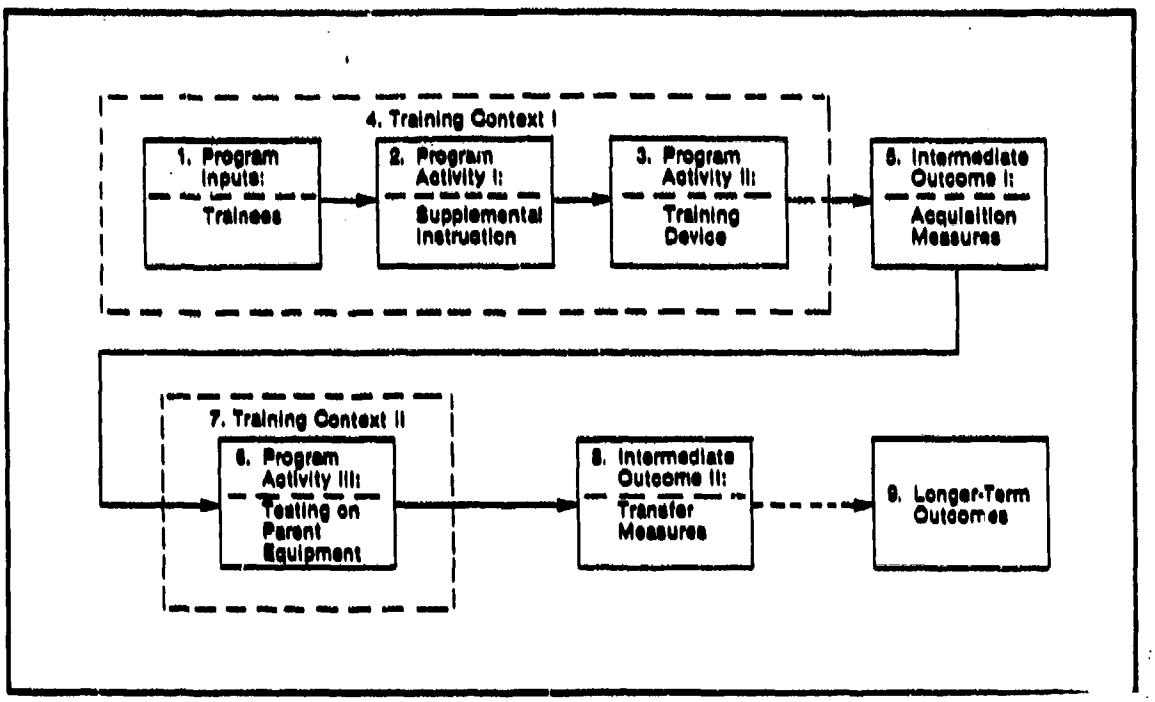
6. **Program activity III** is whatever trainees might do next, such as receiving additional training of some sort or being tested on the parent equipment. In the latter case, we would describe the parent equipment in terms of the tasks comprising the operational performance objectives(s) and its overall similarity to the training device.

7. **Training context II** includes many of the same variables considered under the Training Context I rubric. We are interested in any variables influencing the trainee's interaction with the parent equipment including, for example, instructional features of the training device that are intended to facilitate the interaction, the conditions of performance, the amount of time that has elapsed since cessation of device-mediated training, etc.

8. **Intermediate outcome II** is trainee performance on the parent equipment. This may include measures of initial and later performance as well as several types of process information, all of which may be cast into transfer of training indexes.

9. **Longer-term outcomes** represent the extended effects of the training program. These would include, for example, performance on the parent equipment under wartime conditions, presumably the ultimate criterion of device effectiveness.

The complete program evaluation rationale would be:



We are suggesting that this general program evaluation framework can be used to assess training device effectiveness in terms of the four criterion constructs discussed earlier. There is an <u>acquisition</u> construct representing what is learned on the training device and an <u>acquisition efficiency</u> construct, representing how well (how quickly, cheaply, etc.) the device trains what it is supposed to train. <u>Acquisition of knowledge and skill</u> related to the training objective(s) is measured directly by Intermediate Outcome I, which also provides for assessment of <u>acquisition efficiency</u> in terms of whatever process indexes

are deemed appropriate. At this stage in the evaluation, specific skill acquisition outcomes are interpreted in light of information about trainee, preliminary training, training device and contextual variables.

There also is a transfer construct of device effectiveness, indicating what the trainee will still have to learn after "graduating" from the training device and a transfer efficiency construct reflecting how well the device prepares the trainee for the operational task(s). Both constructs are measured at Intermediate Outcome II by whatever transfer index is judged suitable (e.g., initial transfer, savings, etc.). At this later stage in device evaluation, specific transfer of training outcomes are interpreted in light of information about the degree of overlap between training and operational performance objectives, trainee proficiency on the training device, characteristics of the device and contextual variables.

In essence, the independent and criterion variables that we have described, when considered within a program evaluation framework, define a model of training device effectiveness. A particular training program describes a path between initial inputs, program activities and intermediate outcomes. The distance to the first

intermediate outcome can be expressed in terms of a "deficit" -- how much the trainee must learn in order to attain criterion proficiency on the device, how long it will take him to reach that criterion, and how much it will cost. The distance between the first intermediate outcome (i.e., the acquisition of skill and knowledge on the device) and the second intermediate outcome (i.e., the level of proficiency required on the parent equipment) also can be expressed as a deficit -- how much the graduate trainee still has to learn, how long it will take, etc. Different training devices have different distances or deficits; the four suggested criterion constructs of effectiveness address the magnitude of these distances; the five different classes of independent variables address how rapidly they will be traversed.

The concept of a deficit model of training device effectiveness is depicted in more detail in Figure 2 on the next page. Figure 2 is a stylized representation of various aspects of training devices, the operational task, and the relationships among the several components. Point A represents the initial skills and knowledge possessed by the trainee prior to exposure to the training device or the operational equipment, and the expected level of trainee performance on the operational task prior to

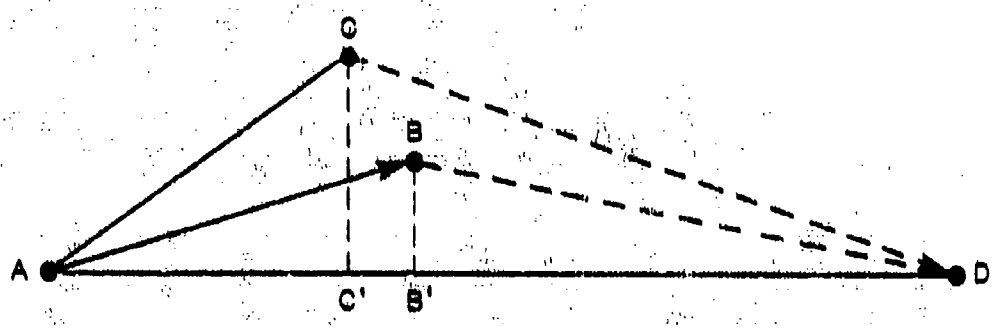


Figure 2. Deficit model of training device effectiveness.

A = initial skills and knowledge of TRAINEE; performance on operational task prior to training on device (TD)

B = skills and knowledge of TRAINEE at completion of TD_1 regimen; criterion performance on TD_1

C = skills and knowledge of TRAINEE at completion of TD_2 regimen; criterion performance on TD_2

D = skills and knowledge needed to perform operational task; criterion performance on operational equipment

B', C' = skills and knowledge needed to perform operational task possessed by trainee after TD exposure; performance on operational equipment

AD = time, cost associated with learning D on operational equipment

AB, AC = time, cost associated with learning B, C on TDs

BD, CD = time, cost associated with learning D given learning on TDs

ABD, ACD = total time, cost associated with learning D for each TD

training. Point D represents the skills and knowledge of performance on the operational task, and the criterion level needed to perform the operational task (using the actual equipment). Thus, the AD "vector" represents a performance deficit and the learning that must occur if the trainee is to learn to perform the operational task. In addition to representing the learning that must take place, this vector also represents the time, cost, and resources necessary to train the operational task <u>using only the operational equipment</u>.

Point B represents the skills and knowledge possessed by the trainee at the completion of training using a training device. It also represents the criterion performance level on the training device, along with the associated time, cost, and resources; the vector BD represents the learning (and associated time, cost, and resources) that is necessary to acquire the appropriate operational skills and knowledge following training on the device. The vector ABD is then the total time, cost, and resources associated with learning D using the training device. Point C and its associated vectors represent a second training device. (This point is included in Figure 2 to allow for situations where alternative training devices are to be compared to each other.) The points B' and C' represent the skills and

knowledge needed to perform the operational task that are possessed by the trainee after exposure to the respective training devices. Hence, B' and C' equate to the trainee's level of performance on the operational task after completion of the training device regimen and prior to any further practice or training on the parent equipment.

The basic rationale for the use of a training device in terms of Figure 1 is that the ABD vector will be "shorter" than the AD vector. That is, the total training cost/time will be less when a training device is used than when the operational equipment itself is used as a trainer.

The ideal training device evaluation, especially when alternative devices or concepts are to be compared, is to measure or estimate ABD and ACD: the total time and cost associated with learning D for each training device, contrasted according to whatever rule the Army may consider appropriate (e.g., cheaper, faster, a cost-time ratio, greater proficiency after a fixed amount of time, etc.).

This evaluation has two major components: an "acquisition" component, conceived as a determination of the time/cost (efficiency) of training to overcome an initial deficit in performance and to reach a criterion level of proficiency on each device; and a "transfer" component,

conceived as an estimation of the remaining trainee deficit that must be overcome in order to demonstrate a criterion level of proficiency on the parent equipment. It is important to keep in mind that the "total" effectiveness of a device is the sum of AB and BD; even if AC is less than AB (i.e., trainees will reach criterion on Device 2 sooner than on Device 1); CD may still be greater than BD (i.e., the remaining deficits are greater Device 2). This could occur, for example, if Device 2 trains all the "easy" parts, while Device 2 trains the "hard" parts. The totals (AB + BD, AC + CD) are not necessarily highly correlated with the acquisition components.

Theoretical Issues: Summary

In this chapter we have discussed a number of theoretical issues related to the evaluation of training device effectiveness. We have described how either an empirical or analytic assessment of effectiveness can be conducted within a program evaluation framework structured around the concept of performance deficits. This approach has the potential of overcoming several limitations found in earlier forecasting models. The performance deficit notion provides a way of operationalizing training importance or criticality considerations. The use of explicit

training program evaluation rationales provides a way of enhancing the diagnostic utility of device evaluation. Finally, the approach we have described broadens the focus of device evaluation to include learning as well as transfer criteria and to permit consideration of the influence of extra-device variables on effectiveness. In the next chapter, we explore some of the real-world constraints on developing and evaluating a training device effectiveness forecasting procedure.

3. Practical and Methodological Issues

In Chapter 1 we traced interest in formal analytic methods for predicting training device effectiveness back to certain constraints associated with the LCSMM and the acquisition process. In Chapter 2 we explored a number of theoretical issues in the course of laying out an analytic approach to device design and evaluation that interrelates a number of predictor and criterion variables within a program evaluation framework. In this chapter we are concerned about practical and methodological constraints on the use and evaluation of the type of forecasting procedures we have been describing. In this connection, three questions are paramount. First, what information is needed to evaluate or estimate device effectiveness? Second, what constraints, if any, does the LCSMM impose on the types and levels of information required to generate predictions of effectiveness? And third, once predictions have been generated, how can we validate them or otherwise assess their quality?

Issue: What Data are Needed to Generate Forecasts?

Assuming that one wants to estimate device effectiveness using the type of analytic procedure just described, then certain information requirements must be satisfied.

Specifically, we need information about the objectives of training and about the independent variables that dictate whether (how well) the objectives will be achieved.

Specification of objectives and variables. Within the context of a training program rationale, it is imperative that the designers and developers of a training device be able to describe the intermediate outcomes they are trying to achieve. Toward that end they need to describe both the operational performance objective for the parent equipment as well as the device-mediated training objective. In spite of the obviousness of this need, and realization that such statements are the sine qua non of any form of device evaluation (i.e., empirical or analytic), it is exceedingly difficult in practice to find adequate specifications. Anyone who seriously doubts this assertion need only review a random sample of Training Device Requirement (TDR) statements to realize how elusive adequate specification really is. As one would expect, the specifications are particularly nebulous during the earlier phases of device acquisition when there is a scarcity of detailed information.

Ideally, specification of the performance objective should be based on operational needs associated with a specific system and one or more missions. When the impetus

for specification of performance objectives comes from the development of a new system, the objectives should properly be defined as an integral part of that system. When the the impetus stems from an observed deficiency in the ongoing performance of some mission-related task, the objectives ought to be specified as part of the "statement of need" that drives the formulation of the training program.

Whatever the impetus for their specification, training and performance objectives can and should be explicitly included in information provided to (or developed by) potential training device/system/program designers and evaluators. They can then be used to derive criterion measures in support of the empirical validation of any actual training approach. More importantly for present purposes, however, they can be used as the starting point for an analytical model to predict the impact of a training device before that device has been actually designed and developed.

As the cornerstones of empirical assessments and analytic evaluations, specifications of performance and training objectives must be defined operationally in such a manner that performance can be reliably and unambiguously measured or otherwise characterized. The operational definition must specify at least the following items:

- the population of subjects to be tested;
- the specific behaviors to be measured;
- the environment for testing (e.g., during daylight); and
- the level of proficiency on the device and/or the parent equipment designated as the criterion.

In the case of Army training, the criterion may be stated as a population statistic, rather than an individual level of proficiency. For example, instead of specifying the performance criterion as some individual score level, the operational criterion may be that 90% of trainees be able to complete a particular task on the training device with no errors. By the same token, specifying the training or performance objective in terms of a single criterion level for each task may be unnecessarily limiting. Instead of a "pass-fail" criterion, it may be preferable to develop a measurement system that discriminates across a range of performance. The latter is desirable, as it permits trade-offs among levels of performance on multiple objectives, and allows aggregation of scores into an overall characterization of performance.

In addition to specifications of training and performance objectives, we need information regarding predictor variables. That is, information about displays, controls, instructional features, task analyses/skill analyses, etc., has to be provided in sufficient detail to be of use to the device analyst/evaluator. In our earlier discussion of forecasting procedures we identified five classes of such variables including trainees, preliminary training, tasks, instructional variables, and the larger training context.

All that we are in fact suggesting in this and the preceding discussion of objectives is that certain data must be available to support analytically derived estimates of training device effectiveness. However, the required data often are not readily available. In the next section, we describe some of the real-world issues that constrain the types and levels of information about training devices and programs.

Issue: How does the LCSMM Affect Device Evaluation?

There have been several recent reviews of training device design and development within the Army system acquisition process (e.g., Kano & Holman, 1982; Matlick, Rosen, & Berger, 1980). In the next few paragraphs, we

will briefly describe the major phases of the training device/simulator acquisition process.

During the first or Evaluation of Alternative System Concepts (EASC) phase, several key decisions are made that ultimately will influence design of the training devices in important ways. For example, based on results of an initial Training Development Study, a Training Device Need Statement is prepared that describes requirements for device-mediated individual and collective training. Alternative training concepts are then considered in the course of selecting a Best Technical Approach to meeting documented needs. These preliminary decisions about the device and its design are reflected in a Concept Formulation Package and an Outline Acquisition Plan.

During the second or Demonstration and Validation (DVAL) phase, the Outline Acquisition Plan is updated and used to acquire an advanced development prototype or breadboard training device. It is during this second phase that the breadboard device is used to support a variety of empirical investigations comprising the Update Training Development Study in which alternative training concepts are assessed and the most promising are validated. The results serve to define the Training Device Requirement and a final Acquisition Plan.

In the third or Full-scale Engineering Development (FSED) phase, the Acquisition Plan is implemented to obtain an engineering development prototype or brassboard training device. At this stage in the acquisition process, design of the training device has been finalized. Production runs are imminent. Assuming that the brassboard device successfully passes various field test evaluations, the fourth or Production phase of acquisition will begin.

The lockstep nature of the training device LCSMM leads to a design dilemma: early on in the device design process, there is very little information available about the parent system upon which design decisions can be based. When such information subsequently <u>does</u> become available, it is usually too late to act on it, to base major design changes in the training device upon it. In other words, while detailed information about the parent system is needed for training system design, design of the device must be initiated before such information materializes in any detail. The consequence of this design dilemma is that the training device design process is a bootstrapping operation, consisting of a series of approximations tied to the evolving structure of the parent system.

As one example of the dilemma, training device designers need, if not detailed descriptions of the parent equipment, at least the job descriptions for system operators. These job descriptions are the source data that serve as input to analytic/rational procedures (e.g., the Instructional Systems Development [ISD] procedures) for determining how best to design and develop training programs. Typically, job descriptions are rendered as Task analyses/Skill analyses (TASA). However, such detailed information, derived from analyses of the parent system, is often too late in coming to be useful in making early and important decisions about training concepts and device design.

Similarly, as we noted in Chapter 1, there are points in the LCSMM where both empirical and analytic evaluations are supposed to occur. For example, the LCSMM provides for an empirical "concept of training" investigation, a "breadboard" evaluation, a "brassboard" evaluation, and Operational Tests I and II. In practice, however, the tight schedule of device development and procurement usually precludes empirical evaluations during the design and development process. Because the training developers have to adhere to the faster-paced materiel system acquisition schedule, time constraints also preclude research on

competing devices or training conceptions early in the acquisition process. If empirical evaluations are conducted (e.g., OT II), they usually occur much too late to modify the device design based on the results. Similarly, while the LCSMM provides for analytic appraisals and review of designs at numerous points, especially during the earlier stages of development, such appraisals, as we noted earlier, are neither systematic nor formalized.

Difficulties in obtaining the right type of information at the proper time are exacerbated by a natural tension between decisions related to instruction and simulation. As a training system matures, it increasingly consists of two environments: an interactive instructional environment, consisting of courseware, adaptive training features, etc., and a simulation environment, consisting of those aspects of the operational situation that are represented in the learning situation. Training developers have to account for the interplay between these two environments during the design and development of a training device. In practice, when one is emphasized, the other is often downplayed, with a potential loss in effectiveness.

Collectively, these and other constraints on information, arising from the realities of the training device LCSMM, have led designers and procurement personnel to exhibit two "tendencies." One is the tendency to gravitate toward high-fidelity devices. This often (but certainly not always) minimizes the "training system" design component. The second is the tendency to adopt a "design to cost" decision rule: design or buy the device with the most instructional features and the highest level of fidelity that is within budget, even though fewer features or lower fidelity may still produce effective training.

Where does all of this leave an analytic model that predicts device effectiveness? The first conclusion to be drawn is that since empirical evaluations of effectiveness are generally infeasible in practice, analytic methods must be used. Second, we believe that sound analytic methods would be used. Designers and developers are forced by circumstances beyond their control to make analytic assessments, but have few if any analytic tools with which to work. Good methods would rapidly find their way to the appropriate audience. Finally, these methods must be flexible enough to allow evaluations to occur with a wide range of input information -- from very general "training

concept" speculations early in device acquisition to very detailed engineering specifications later on. The challenge is to conceive of ways in which estimates of effectiveness can be generated that overcome the many constraints we have alluded to.

Issue: How Can Forecasts be Validated?

How would one go about determining the validity of a device effectiveness forecasting model? An obvious suggestion is to use empirical data. It is unfortunate in this regard that opportunities to try out analytic models and to use the results of empirical tests to revise the models for improved prediction have been extremely limited. Tryout and revision would require reliable measurement of both predictors and criteria. Practical constraints (cost; limited availability of devices, parent equipment, trainees, and subject matter experts) have limited the cases in which both criterion and predictor measurement were reported (e.g., Wheaton & Mirabella, 1972; Mirabella & Wheaton, 1973; Wheaton, Rose, Fingerman, & Leonard, 1976c).

Part of the measurement infeasibility problem derives from the explicit assumption of many analytic procedures that they should be predicting transfer to operational equipment as the index of device effectiveness. Hence,

major components of these models (e.g., "Commonality," "Similarity," etc.) are structured around comparisons between a training device and the operational equipment. It follows that any evaluation or testing of such models must use parent equipment performance as the criterion.

However, even when criterion measures are defined more broadly to include acquisition phenomena and when arrangements can be made to collect predictor and criterion data, other problems persist. The most fundamental of these is that validation of forecasting procedures, or research on the component variables and weightings underlying such procedures, invariably requires some form of regression paradigm.

Regression paradigms in which device features are systematically varied and then related to obtained (empirical) effectiveness scores are at best infeasible. Since the number of variations in device or training program features is probably greater than the number of devices, one would not have enough degrees of freedom to conduct a regression analysis. Furthermore, there usually are not sufficient numbers of alternate devices that will have been produced to allow for significant variability in any criterion

measures of effectiveness.[4]

To illustrate this problem, consider a hypothetical training system evaluation effort: several devices are used. Predictions of effectiveness are generated for each device. Then the devices are used in training and transfer experiments and actual results are compared to predicted values.

What we might find is that Device A, with high-fidelity stimuli, motion cues, moderate response similarity, no augmented feedback, and no freeze-frame capability did slightly better than Device B, which contained low-fidelity stimuli, motion cues, high response similarity, augmented feedback, and no freeze-frame capability, which did much better than Device C with . . . . Clearly, we have little hope of untangling these outcomes to determine the critical device dimensions contributing to different levels of effectiveness. Are there other approaches to evaluating and refining forecasting models?

---

[4] A possible approach to this problem of insufficient numbers of alternative devices is being investigated by ARI. This approach involves laboratory experiments with "real" training devices, where the experimenter artificially creates several versions of the same device, trains groups of subjects on each version, and "transfers" all of the subjects to a single "criterion" version.

Alternative empirical approaches. A different approach to measuring effectiveness is contained in the program evaluation approach described in the preceding chapter. The concept is that if "ultimate" objectives cannot be measured, the intermediate objectives and the links between the various objectives can be. For example, it may be relatively easier to measure acquisition performance on the training device. These scores could be used as criterion data for assessment of program features, such as individual difference variables, user acceptance indexes, etc.

Again, assuming that it is not possible to measure transfer to the operational system, we may still be able to generate indirect or inductive support for device effectiveness. The argument is as follows: Transfer to a specific operational task is, in essence, a generalization phenomenon: Will good performance in one set of circumstances generalize to other circumstances (of which the parent equipment is only one example)? That is, will performance be maintained with a variety of stimuli, a variety of responses, different controls, different environmental circumstances, etc.? Evidence of generalization can be used as inductive evidence for transfer to a particular (i.e., operational) situation.

Thus, one could use a series of surrogate transfer/generalization situations, perhaps including different training device configurations and other analogous equipment, to test the generalizability of acquired skill and knowledge. Our confidence in the effectiveness of a device would increase with each demonstration of generalization to a different device configuration.

In conjunction with alternative empirical approaches, the program evaluation framework prescribes certain analytic and statistical methods that can be used to validate a device effectiveness forecast model. Specifically, when any analytic method is used to generate predictions of training effectiveness, a number or set of numbers is produced. Is there anything that can be done with these numbers to determine their potential usefulness without collecting actual performance data? In the following sections, we describe several analyses that directly or indirectly may shed light on the validity of any proposed forecasting procedure.

Sensitivity analyses. Suppose we generate a set of numbers meant to represent the effectiveness of two devices. For example, Device 1 is estimated to have an effectiveness of 0.20 and Device 2 is estimated at 0.25.

Is the difference between 0.20 and 0.25 "significant," i.e., would we expect soldiers trained on one device to perform better than soldiers trained on the other? Or is this difference within the measurement error of the estimation system? To answer these questions, it is necessary to derive a distribution for any predictive index that allows statements about differences in predicted values.

One very interesting question is "sensitivity": whether or not a set of ratings differs significantly from that which would be obtained by random assignment of ratings to the available scales. With a lack of knowledge about distributional characteristics of model parameters, the assumption of uniform distributions provides the most diffuse values. Investigation of this problem also pinpoints some of the problems that will surface in investigating other potential distributions.

Reliability. The reliability of an estimate of effectiveness is determined by the reliabilities of its constitutents. That is, once the reliabilities of the operational measures of variables are determined, the reliability of a measure of effectiveness (which is a combination of operational measures) may be calculated. For simple combination rules, it may be possible to determine

analytically the reliability of the combined measure. For other, more complex combinatorial rules, it may be more reasonable to determine the reliability by Monte Carlo simulation.

One of the most important analyses that can take place in the evaluation of estimates of effectiveness is the examination of the properties of the rules, to determine whether they are sensible and whether they predict desired properties of an effectiveness measure. For example, if effectiveness is a multiplicative combination of the constituent variables, one would expect there to be a zero point for each constituent such that effectiveness would be a constant whenever at least one of the constituent measures was at the zero point. On the other hand, additive rules do not have this property. The properties of any effectiveness measure that is a simple polynomial can be examined by looking at its additive and multiplicative components. In addition, properties of the combination rules at the extremes will give an indication of the validity of the rules.

Incremental validity. One standard method for assessing validity is to compare the predictions of the combination rules to expert judgments. The methods of

conjoint measurement, policy capturing (using multiple regression), and functional measurement (using analysis of variance) can be applied to compare expert judgments with the predictions of the model. These three methods differ in basing their tests either on ordinal or on interval properties of the data, and in requiring or not requiring a balanced design. This evaluation uses expert judges to define the reasonableness of combination rules, and it performs an analysis similar in many ways to the logical analysis of properties described above.

The analysis of the history of devices for which longitudinal archival data were available would give a further indication of the validity of the estimate of effectiveness. For example, we would expect that the effectiveness of a device would increase as it was modified and improved, and as problems with it were fixed. Thus we would expect our prediction of effectiveness to mimic the notions of device effectiveness that were being used by the decision makers. If it did, this would argue for the validity of our predictive estimate. In other words, if the predicted score increased as the device became more highly developed, we would expect the validity of the estimate to be strengthened.

There is another way that we may obtain information relevant to the validity of the estimate of effectiveness, again from an historical analysis of decisions made during the development of the device: Basically, at any stage in the process, development of a device may be continued or it may be stopped. At earlier stages in the acquisition cycle, development of a device may continue either if the design is promising or to obtain more information regarding its estimated effectiveness. It would be expected that at any stage, the decision to continue -- that is, the decision to "purchase" more information about the device -- would be related to the measurement of effectiveness. As was pointed out above, the validity of the predicted estimate of effectiveness would be expected to increase for devices in later stages of development. If we assume that the decision maker is (or should be) considering this, we can compare our estimate to the history of these decisions. Ultimately, it may be possible to model these information-purchasing decisions to aid the decision maker further.

Discriminability. The discriminability of an aggregate measure of effectiveness depends on the aggregation rule and on the joint distribution of values of the individual constituents of the effectiveness measure. For example, if the combination rule is additive and

constituents are, in general, negatively correlated, the aggregate measure will not discriminate among devices. Consequently, the weights that are used in the effectiveness model will have a great effect on the relative measures of the effectiveness of two devices. Since negative correlations may be the product of the tradeoffs that the designer of the device makes to arrive at a product with a reasonable cost, it is likely that the effectiveness scale will have low discriminability.

One way to investigate the discriminability of the measure is to compare actual devices known to differ in effectiveness. This comparison gives an indication of the ability of the measure to detect large differences in effectiveness. Another way to investigate the discriminability of the predicted effectiveness measure is to conduct Monte Carlo simulations in which hypothetical devices are evaluated. The distributions of the scores on the constituent variables are varied; for some cases, the variables positively correlated; for others, the variables independent or negatively correlated. Finally, distributional properties of the overall measures can be examined.

Efficiency. The best measure of "effort" in determining the efficiency of a measure is the number of constituent variables that make up the aggregate measure. The actual form of the combination rule is probably unimportant in assessing effort. Thus, validity/number of constituents is a reasonable measurement of efficiency in this measure, just as error-reduction/degrees of freedom is a reasonable method of testing models in the analysis of variance. In this sense, efficiency is a measure of the parsimony of the model. A measure of efficiency which includes a large number of variables requires great "effort" and is unparsimonious.

Simplicity. The lack of an effectiveness criterion requires in most cases that the model with the most parameters be taken as the criterion. A critical question to ask is whether some smaller set (which presumably could be more reliably and efficiently obtained) could produce the same predictions. This would obviate the necessity for cumbersome and potentially unreliable calculations and judgments. If we consider the predictions of the most complex model as a criterion, we could use stepwise regression techniques to determine the relative ability of simpler models to give the same results as the most complex model. In addition, using standard statistical tests, we could

compare different (and perhaps simpler) functional forms for the effectiveness measure with the most complex (and presumably most accurate) measure. For example, the ratio of goodness-of-fit measures could be compared using an F-test.

Care should be taken, however, in considering these simplicity analyses. While simplicity is an important virture for this particular use of the model (i.e., generating a single measure of "predicted effectiveness"), it may not be desirable for other uses of the model, such as diagnostic power.

Practical and Methodological Issues: Summary

To be maximally useful, any model must be sensitive to variations in the quality and quantity of input information. For decisions early in the LCSMM, not much more than general "function" statements are available regarding task and training demands. There are insufficient data to conduct all but the most general types of analyses and to make only the grossest of decisions regarding training device (or system) concepts. As more data become available -- both about the operational task and equipment, and about the proposed training system -- more detailed judgments and estimates of effectiveness can be made.

Thus, the practical constraints of the LCSMM require that an effectiveness evaluation model be capable of generating predictions with both general and detailed inputs. Similarly, and perhaps more importantly, models should be capable of providing diagnostic information -- why the design concept is judged ineffective, how a design concept could be improved -- at all stages of development.

There also are practical constraints on the evaluation of a device effectiveness forecasting system. One approach is to conduct the required empirical tests when feasible. When infeasible, other less direct assessments may be required. These must be designed to accumulate presumptive evidence for the validity of the forecasting models. It is essential that development and evaluation of these models continue, despite these practical obstacles. In this chapter, we have suggested several directions in which to proceed.

REFERENCES

Aagard, J.A., & Braby, R. (1976). **Learning guidelines and algorithms for types of training objectives.** TAEG Report No. 23. Orlando, FL: Training Analysis and Evaluation Group.

Blaiwes, A.S., Puig, J.A., & Regan, J.J. (1973). **Transfer of training and the measurement of training effectiveness.** **Human Factors, 15,** 523-33.

Blaiwes, A.S., & Regan, J.J. (1970). **An integrated approach to the study of learning, retention, and transfer -- a key issue in training device research and development.** NAVTRADEVCEN IH-178 (AD 712096). Orlando, FL: Naval Training Device Center.

Braby, R., Henry, J.M., Parris, W.F., Jr., & Swope, W.M. (1975). **A technique for choosing cost-effective instructional delivery systems** (TAEG Rep. No. 16). Orlando, FL: Department of the Navy, Training Analysis and Evaluation Group.

Carroll, R.M., Rhode, A.S., Skinner, B.B., Mulline, J.L., Friedman, F.L., & Franco, M.M. (1980). **Manpower, personnel, and training requirements for materiel system acquisition** (Draft). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

CORADCOM. (1980). **Training acquisition handbook.** Fort Monmouth, NJ: U.S. Army Communications Research and Development Command.

Cronin, R.C., & Bourque, B.B. (1981). **National Evaluation Program Phase I Report: Assessment of victim/witness assistance projects.** Washington, DC: U.S. Department of Justice, National Institute of Justice.

Cronin, R.C., Drury, M.J., & Gragg, F.E. (1983). An evaluation of the FmHA-AoA demonstration program of congregate housing in rural areas. Final Report. Washington, D.C.: American Institutes for Research.

Faust, D.G., Sweezey, R.W., & Unger, K.W. (1984). Field application of TRAINVICE. A study of four models designed to predict training device transfer-of-training potential, Draft. Alexandria, VA: U.S. Army Research Institute.

Gagne, R.M. (Ed.). (1963). Psychological principles in system development. New York, NY: Holt, Rinehart, and Winston.

Gagne, R.M., Foster, H., & Crowley, M.E. (1948). The measurement of transfer of training. Psychological Bulletin, 45, 97-130.

Gibson, J.J., & Gibson, E.G. (1955). Perceptual learning: Differentiation or enrichment? Psychological Review, 62: 32-41.

Harris, J.H., & Ford, P. (1983). Application of transfer forecast methods to armor training devices. FR-TRD(VA)-83-3. Alexandria, VA: HumRRO.

Hammerton, M. (1963). Transfer of training from a simulated to a real control situation. Journal of Experimental Psychology, 66, 450-453.

Hays, R.T. (1980). Simulator fidelity: A concept paper, Technical Report 490. Alexandria, VA: Army Research Institute.

Hirshfeld, S., & Kochevar, J. (1979). Training device requirements documents guide. Orlando, FL: PM TRADE, Naval Training Equipment Center.

Kane, J.J. (1981). Personnel and training subsystem integration in an armor system, Research Report 1303. Fort Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences.

Kane, J.J., & Holman, G.L. (1982). Training device development: Training effectiveness in the Army system acquisition process. McLean, VA: SAI.

Kinton, Inc. (1980). **Sources of information on integrated personnel and training planning: A handbook for TRADOC system managers (TSM)**, MESTA 80-1. Alexandria, VA.

Knerr, C.M., Nadler, L., & Dowell, S.K. (1983). Comparison of training transfer and effectiveness models. **Proceedings of the Human Factors Society 27th Annual Meeting.** San Diego, CA: Human Factors Society.

Mackie, R.R., Kelly, G.R., Moe, G.L., & Mecherikoff, M. (1972). **Factors leading to the acceptance or rejection of training devices.** Orlando, FL: NAVTRAEQUIPCEN.

Matlick, R.K., Rosen, M.H., & Berger, D.C. (1980). **Cost and training effectiveness analysis performance guide.** Springfield, VA: Litton Mellonics.

Miller, R.B. (1954). **Psychological considerations for the design of training equipment.** Pittsburgh, PA. American Institutes for Research.

Mirabella, A., & Wheaton, G.R. (1973). **Effects of task index variation on transfer of training criteria,** Technical Report. NAVTRAEQUIPCEN 72-C-012601. Orlando, FL: U.S. Naval Training Equipment Center.

Murdock, B.B., Jr. (1957). Transfer designs and formulas. **Psychological Bulletin, 54**(4), 313-325.

Narva, M.A. (1979a). **Formative utilization of a model for the prediction of the effectiveness of training devices,** Research Memorandum 79-6. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Narva, M.A. (1979b). **Development of a systematic methodology for the application of judgmental data to the assessment of training device concepts,** Research Memorandum 79-7. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Neisser, U. (1976). **Cognition and reality. Principles and implications of cognitive psychology.** San Francisco, CA: W.H. Freeman and Company.

Osgood, C.E. (1949). The similarity paradox in human learning: A resolution. **Psychological Review, 56,** 132-143.

Roscoe, S.N. (1971). Incremental transfer effectiveness. **Human Factors**, 13(6), 561-567.

Roscoe, S.N. (1972). A little more on incremental transfer effectiveness. **Human Factors**, 14, 363-364.

Rose, A.M., Allen, T.W., & Johnson, E.J. III. (1982). **Acquisition and retention of soldiering skills: Development of a task classification system.** Washington, DC: American Institutes for Research.

Rose, A.M., McLaughlin, D.H., & Felker, D.B. (1981). **Retention of soldiering skills: Review of recent ARI research.** Washington, DC: American Institutes for Research.

Rose, A.M. (1980). Information-processing abilities. In R.E. Snow, P.A. Frederico, and W.E. Montague (Eds.), **Aptitude, learning and instruction. Volume 1: Cognitive process analyses of aptitude.** Hillsdale, NJ: Lawrence Erlbaum Associates.

Royer, J.M. (1979). Theories of the transfer of learning. **Educational Psychologist**, 14, 53-69.

Smode, A.F. (1972). **Training device design: Human factors requirements in the technical approach,** Technical Report, NAVTRAEQUIPCEN 71-C-0013. Orlando, FL: Dunlap & Associates.

Swezey, R.W., Criswell, E.L., Huggins, R.S., Hays, R.T., & Allen, J.A. (1985). **Training effectsof hands-on practice and of three instructional delivery methods on maintenance performance in a simple engine repair procedure.** Draft Report. Alexandria, VA: U.S. Army Research Institute.

Swezey, R.W., & Evans, R.A. (1980). **Guidebook for users of TRAINVICE II.** Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Thorndike, E.L. (1903). **Educational psychology.** New York, NY: Lemke and Buechner.

Thorndike, E.L., & Woodworth, R.L. (1901). The influence of improvement in one mental function upon the efficiency of other functions. **Psychology Review**, VII, 247-261, 384-395, 556-564.

Tufano, D., & Evans, R. (1982). **The prediction of training device effectiveness: A review of army models.** ARI Technical Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Valverde, H.H. (1968). **Flight simulators: A review of the research and development.** Dayton, OH: Wright-Patterson Air Force Base, Aerospace Medical Research Laboratory.

Wheaton, G.R., Fingerman, P.W., Rose, A.M., & Leonard, R.L., Jr. (1976a). **Evaluation of the effectiveness of training devices: Elaboration and application of the predictive model,** Research Memorandum 76-16. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wheaton, G.R., Rose, A.M., Fingerman, P.W., Korotkin, A.L., & Holding, D.H. (1976b). **Evaluation of the effectiveness of training devices: Literature review and preliminary model,** Research Memorandum 76-6. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wheaton, G.R., Rose, A.M., Fingerman, P.W., & Leonard, R.L. (1976c). **Evaluation of the effectiveness of training devices: Validation of the predictive model,** Final Report. Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Wheaton, G.R., & Mirabella, A. (1972). **Effects of task index variations on training effectivenss criteria,** Final Report. NAVTRAEQUIPCEN 71-C-0059-1. Orlando, FL: U.S. Naval Training Equipment Center.

APPENDIX A

Appendix A: Indexes of Transfer and Theoretical Bases

There are several commonly used indexes of transfer. For example, it is possible to express the amount of transfer between a training device and the parent operational equipment relative to the performance of an untrained control group of soldiers on the parent equipment (e.g., Gagne, Foster, & Crowley, 1948):

Percentage of Transfer = [(E - C) / C] X 100.

In this formulation, E refers to the performance of the experimental group of soldiers on the parent equipment following training on the training device, and C refers to the performance of the control group of soldiers on the parent equipment, not having been trained on the training device.

Another commonly used index is to compare the obtained transfer with the "maximum possible value" (Murdock, 1957). The maximum possible value is the best score hypothetically attainable on the parent equipment:

Percentage of Transfer = [(E - C) / (T - C)] X 100,

where T is the maximum possible score.

A third index expresses transfer as the ratio of the difference between the experimental and control scores to the sum of these scores (e.g., Murdock, 1957):

Percentage of Transfer = [(E - C) / (E + C)] X 100.

All of the above formulations can be applied equally well to first-trial or "cumulative" (i.e., summative) performance. However, more elaborate indexes of transfer are necessary when learning rates are considered (e.g., Roscoe, 1971; 1972). The skill acquisition curve for the operational task on the parent equipment must be described by at least two parameters: the performance level at the beginning of practice (i.e., "initial transfer") and the rate of change in performance across practice. It is entirely possible that different characterizations of device effectiveness might be associated with these two parameters. For example, Hammerton (1963), using an airplane simulator, found initial negative transfer, but positive long-term transfer (i.e., total "savings" on time to criterion on the operational task).

Just as there are several popular empirical indexes of transfer, there also are different perspectives about its theoretical underpinnings. The theoretical bases of the transfer phenomenon have a long history in applied

psychology, dating back to Thorndike (e.g., Thorndike & Woodsworth, 1901; Thorndike, 1903). He proposed a theory of "identical elements," claiming that there would be positive transfer in the learning of a second task to the extent that that task required components learned in some other task. In this view, transfer was quite specific. Facilitation of performance on the new task would not occur unless at least part of the new task consisted of "elements" specifically learned in the first task.

More commonly, transfer is formulated in stimulus-response terminology, with the Osgood (1949) transfer surface as the principal exemplar: the amount and direction of transfer vary as a function of stimulus and response similarity between two tasks. According to the Osgood surface, when the stimuli for two tasks are identical but the responses are completely unrelated, maximum negative transfer theoretically will occur. Maximum positive transfer is expected when both stimuli and responses are identical for the two tasks.

In current cognitive psychological terminology, transfer depends on the modification of pre-existing knowledge structures ("schemas") by training so that new information (e.g., about the next task to be learned) can be

efficiently incorporated (e.g., Neisser, 1976). Transfer will occur when, during practice on an initial task, new information is added to existing knowledge bases that trainees can apply to the second or new task.

We also can consider the transfer paradigm as a strategy selection situation (e.g., Gibson & Gibson, 1955). When faced with a new task, people apply previously learned strategies. The selection of a particular strategy depends upon the perceived degree of similarity between the new situation and whatever the performer has previously learned. If the circumstances or context of the new task is similar to that of the previously learned task, trainees will try the strategies that were previously successful. Postive transfer will occur if these strategies are "appropriate"; no transfer or even negative transfer will occur if the perceived similarity leads to the selection of <u>inappropriate</u> strategies -- that is, the trainee perceives (and acts on) a similarity when none exists.